ARCH 3819 Text mining intro

This page is a companion to a guest lecture prepared for ARCH 3819/ARCH5819.

Class Preparation

Please bring a laptop. If you do not own one, feel free to check one out at the Olin circulation desk. Our explorations are intended to be participatory.
No special software will be needed. All explorations will be done through a Web browser.

Discussion of terms

There will be a short discussion where we define terms (at least loosely) so that we might use a common language for our discussion and critique of these tools and their strategies.

Explorations

All analysis tools will be demonstrated; no prior knowledge of any tools will be required. The room is equipped with video input to allow sharing of your desktop on the screen. If you discover something that you would like to share and discuss, we can easily do so.

Word Frequencies

Voyant is a low barrier text analysis tool that delivers an interactive interface and a variety of visualizations.

- **Data**: Text of your choosing. Upload interface accepts files in the following formats: plain text, PDF that has embedded OCR, MS Word doc and docx files, URLs. Can take multiple documents to make a set. Upload of any material will be subject to the Voyant privacy policy.
- **Analysis**: Voyant environment lists every word in the document and counts for each. Also calculates frequencies based on the total word word count of the document(s) in the uploaded set.
- **Visualization**: Interactive dashboard: word cloud, graph of frequency over document segments, tabled secondary data, source data as text. Also has navigational aids that integrate source an secondary data with each other and with visualizations, and offers utilities for stop word control, URLs, screen capture, secondary data download, etc.

Sample texts and URLs for analysis are listed below for experimentation, but feel free to use other source data that interests you.

- Sample texts, courtesy of Project Gutenberg. Use the **plain text** version.
    - Crane, Stephen, 1871-1900. The Red Badge of Courage: An Episode of the American Civil War.
    - Dickens, Charles, 1812-1870. A Tale of Two Cities.
    - Upham, Charles Wentworth, 1802-1875. Salem Witchcraft, Volumes I and II
- Sample URLS: copy and paste into the Voyant upload browser window to get started.
    - DSPS Press - http://blogs.cornell.edu/dsps/
    - The Dirt - http://dirt.asla.org/
    - The Transport Politic - http://www.thetransportpolitic.com/
    - City Of Sound - http://www.cityofsound.com/

nGrams

nGrams depict the frequency of a word or word phrase and are most often depicted over publication year. We have two nGram tools, each leveraging different source data.

Examples from Quantitative Analysis of Culture Using Millions of Digitized Books. Jean-Baptiste Michel, et. al *Science:* 14 Jan 2011:Vol. 331, Issue 6014, pp. 176-182; DOI: 10.1126/science.1199644

Google's nGram Viewer. The links below as starting points. Dynamic modifications can be made at any point. Rules for syntax can be found on the About page.

- **Data**: Primary source data was from selected volumes from Google Books (methodology is described here).
- **Analysis**: Secondary data of ngrams derived through algorithms that tokenize text and then count the frequencies of those tokens. Secondary data of tokens, counts, frequencies, and publication dates. Tool is integrated with secondary data.
- **Visualization**: Simple line graph plots of ngram frequency over publication date from secondary data.
- Examples
    - Urban renewal
    - Greek Revival,Gothic Revival,Queen Anne Style
    - Nicknames for cities
    - Spanish words in English publications

    - Frequency for which Roland Barthes has been cited in various languages
    - "Ebola" 1975-2008 in various languages (compare with data from WHO on Ebola outbreaks and a map of official languages of African countries)

HathiTrust Research Center (HTRC) Bookworm

- **Data**: Primary source data is the openly viewable texts from the HathiTrust Research Center (publications mostly pre-1923).
- **Analysis**: HTRC algorithms tokenize text and then count the frequencies of those tokens. Secondary data includes tokens, counts, frequencies, publication dates, etc. arranged in SOLR indexes.
- **Visualization**: Simple line graph plots of ngram frequency over publication date of the secondary data. Tool has faceting control as well that leverage enriched bibliographic metadata of the texts comprising the primary data.
- Examples
    - Spanish Words in English publications
    - Various words for "creole"
    - Mystery (contrast fiction and nonfiction)

Network Analysis

Example: **Linked Jazz** is an interactive visualization of various types of connections between notable Jazz artists.

- **Data**: Transcripts of oral histories from the archives of several institutions. (See https://linkedjazz.org/data-sources/ for explanation.)
- **Analysis**: Entities (names of Jazz artists) are extracted, and relationships analyzed by both automated and human (crowd-sourced) means. (See https://linkedjazz.org/data-productionworkflow-draft/ for details.) RDF triples (linked data) created.
- **Visualization**: Nodes and connecting edges that represent entities and relationships between them.

Immersion is a tool for analyzing and depicting connections in email. By design, Immersion collects only header information (From, To, Cc and Timestamp). The FAQ describes what information you grant access to, how it will be used, and how to delete your data when you are done. You can also explore with demo data.

- **Data**: Gmail, Yahoo or MS Exchange (unhosted only) mail account.
- **Analysis**: Done in real time. Immersion parses mail headers for flow information.
- **Visualization**: Displays entities (senders, receivers, etc.) as nodes and sending/receiving connections as lines. The slider at the foot of the visualization controls for time. Size of nodes pertains to volume of mail sent.

Gephi is a commonly used network analysis tool that is much more flexible and powerful. It requires local installation.

Spatial and Temporal Representation

Viewshare is a free tool provided by the Library of Congress that generates interactive maps and timelines with facets for digital collections. The tool presupposes the setup of a user account and data in columnar form that includes location and/or time related data fields. A few helpful tutorials are available.

- **Data**: User supplied. The example below is from selected metadata from the Cornell HipHop Collection. It included Photographs from Joe Conzo of the early HipHop Music scene and flyers of HipHop venues.
- **Analysis**: The photos and flyers were cataloged in Shared Shelf with spatial and temporal information of the subjects. Resulting metadata was downloaded in MS Excel format.
- **Visualization**: Spreadsheet was uploaded into Viewshare, ans minimally set up. Map and timeline tie to uploaded secondary data.

Test visualization of digital collections from Cornell University Hip Hop Collection created with Viewshare

Image Analysis

Ukiyo-e.org is a database and image similarity analysis engine, created by John Resig to aide researchers in the study of Japanese woodblock prints.

- **Data**: Over 213,000 digital copies of prints from 24 institutions, and their cataloging metadata. Metadata is indexed and searchable. Details are noted in the about page.
- **Analysis**: Image search uses the TinEye matching engine to determine edges in an uploaded sample and compares with analyzed edges in database, returning probable matches.
- **Visualization**: Tiled images of "hits" for easy comparison, with URL links to their metadata in the source institution's catalog.

Sample Images for search - click on desired image to display and choose either "download" or "save as..."